UPMC
SORBONNE UNIVERSITÉS

ALGORITHMES SUR LES ARBRES ET LES GRAPHES EN BIOINFORMATIQUE

Résumé des articles

1- Next generation proteomics: Towards an integrative view of proteome dynamics.

2- Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.

Elaboré par : **WANG Jinxin**
Master1 BIM groupe: **.....**
Année universitaire : **2014/2015**

Résumé de l'article 1

Altelaar AF, Munoz J, Heck AJ. 2013. Next generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics.* January 2013, pp. 35-48.

La nouvelle génération d'analyse protéomique: vers une vue intégrative de la dynamique des protéomes

Le génome humain contient 20,300 gènes. L'expression spécifique d'une partie de ce génomes (~11,000 gènes) détermine le squelette moléculaire du phénotype cellulaire. Les mécanismes complexes de régulation protéique telle que la maturation, la modifications post-traductionnelles (PTMs), les interactions protéine-protéine (PPIs) et la localisation subcellulaire induisent une dépendance temporelle du réseau protéique tissu-spécifique dont les réponses aux perturbations sont variable.

Cet article expose les récentes stratégies qui on permis le développement de la nouvelle génération d'analyse protéomique par spectrométrie de masse (MS- based proteomics), les récentes applications et les progrès technologiques de la protéomique ainsi que les applications de l'expression quantitative des protéomes dont les données sont complémentaires à l'information portée au niveau des gènes et des transcrits. Ceci couvre la diversité des PTMs avec un regard sur la PPI et les réseaux de signalisation avec des réponses dynamiques aux perturbations. La revue conclut par une discussion des applications cliniques de la protéomique basée sur la spéctrometrie de masse vers une vue intégrative de la dynamique du protéome dont la variation temporelle et spatiale est dépendante de la réponse de l'organisme à son environnement.

La maturation de l'ARNm et les modifications post-traductionnelles des protéines induisent une grande variabilité de protéines traduite à partir d'un gène unique. La taille et la complexité du protéome est ainsi plus importante que celle du génome. La technologie actuelle d'analyse protéomique (‘next-generation proteomics’) reflète cette habilitée à caractériser des protéomes entiers.

Afin d'atteindre un degré optimal d'identification protéomique et une meilleure couverture de la quantification, une implémentation de plusieurs étapes est nécessaires. Les récentes avancées technologiques et méthodologiques ont augmenté considérablement la portée de la couverture protéomique, ainsi on peut citer: la fluorescence-activated cell sorting (FACS) ou tissue micro-dissection qui permet de cibler des échantillons de plus en plus complexes à partir de quantité encore plus faible d'échantillons, mais aussi la hydrophilic interaction liquid chromatography (HILIC) . La chromatographie en phase inverse directement couplée à l'analyse par spectrométrie de masse permet une meilleure réduction de la complexité. Les approches actuelles utilisent la chromatographie en phase inverse C18 sous un nanoflux et des régimes de haute pression. Finalement la quantité d'information extraite peut être augmentée par une amélioration de l'efficacité d'identification en utilisant une haute résolution de la masse et des technologies de fragmentation complémentaires à ion-trap collision-induced

dissociation (CID) tel que electron transfer dissociation (ETD) et higher-energy collisional activation (HCD).

L'ultime objectif de la protéomique quantitative est la détection parallèle de différents échantillons après plusieurs temps de perturbation comprenant réplicat technologique et biologiques qui procurent une puissance statistique avec un temps d'expérience minimal.

L'analyse protéomique par spectrométrie de masse:

Dans la vaste majorité des expériences proteomiques, les protéines extraites sont digérée en peptides par la trypsine pour créer des espèces moléculaires faciles à manipuler dans l'analyse par spectrométrie de masse. La digestion enzymatique de tout le protéome génère des centaines de milliers de peptides: cet échantillon complexe ne se trouve pas directement compatible avec l'analyse MS. Cependant, le premier pas du travail proteomique est souvent dirigé par la réduction de la complexité de l'échantillon. Dans le cas du préfractionnement de l'échantillon, la population de peptides est fractionnée selon les propriétés physicochimiques telle que la charge, le point isoelectrique, l'hydrophobicité ou une combinaison de ces propriétés. Il est essentiel que la technique de préfractionnement choisie soit orthogonale à la séparation de la chromatographie en phase liquide (LC) juste avant l'analyse par MS. Alternativement, des sous ensembles de l'échantillon peuvent être ciblés à travers l'enrichissement de peptides contenant des modifications (phosphorylation (P), dimethylation (Me2) ou acetylation (Ac) par exemple) en utilisant des résines basées sur l'affinité ou une immunoprecipitation par anticorps (IP). Ces échantillons préfractionnés ou enrichis sont ensuite introduit dans le système LC pour une étape de séparation supplémentaire afin de réduire davantage la complexité; ceci peut être accompli en utilisant la chromatographie en phase liquide à haute performance ou ultra-high-performance liquid chromatography(UHPLC). Après ionisation, les ions précurseur peptidiques sont introduits dans le spectromètre de masse, qui enregistre leur ratio de masse sur charge (m/z) avec une grande précision. Pour l'identification, chaque précurseur est sélectionné (sur la base de l'intensité observée) pour une analyse de MS en tandem (MS/MS)- plus communément, collision-induced dissociation (CID) —pour générer des fragments d'ions caractéristiques pour le précurseur sélectionné. La combinaison de m/z du précurseur et ses fragments d'ions est alors corrélée pour connaitre la séquence peptidique à partir de plus grande bases de données protéiques en utilisant des algorithmes tel que as Mascotor SEQUEST. Les données sont ensuite quantifiées d'une manière relative ou absolue.

quantification proteique sans marqueurs:

La quantification protéique sans marqueurs à travers le comptage spectral et/ou l'intensité du signal des peptides détectés est utilisée afin d'obtenir des informations quantitatives de l'expression protéique relative si effectuée par mesures de réplicas sous des conditions rigoureuses avec une variation d'analyse minimale.

Selected reaction monitoring (SRM) est une stratégie émergente de quantification qui cible des protéines d'intérêt spécifiques , utile pour la validations des changements d'expression protéiques. Les

essai SRM atteignent un haut degré de spécificité à travers la surveillance de l'unique combinaison du ratio de la masse sur charge du peptide et de multiples ions de fragments peptidiques.

quantification proteique avec marqueurs:

Différentes stratégies ont été développées basées sur l'incorporation d'isotopes stables, dont une variabilité minimale peut est atteinte en utilisant le marquage métabolique dans des cultures cellulaires ou dans un organisme entier. La quantification est généralement effectuée au niveau du MS, à l'exception du marquage chimique basé sur le marquage isobarique, dans lequel la quantification est basée sur les ions rapporteur du MS/MS. La quantification au niveau de MS/MS peut être "muliplexée" (multiplexed (up to octoplex)), permettant de ce fait l'analyse des multiples perturbations en parallèle.

Elaboration du profil d'expression du proteome

La cartographie de l'ensemble de protéines dans un système biologique et la compréhension du réajustement spatio-temporel des protéines représente l'un des principaux objectifs de la protéomique.

Les avancées considérables du séquençage à haut débit de l'ARN (RNA-seq) ne permettent pas de connaitre le nombre de gènes codent des protéines. Cette incertitude est partiellement due au fait que le génome est extensivement transcrit en ARN non-codant, pour lequel l'importance biologique reste à explorer. La protéogénomique se défini comme étant l'utilisation de l'information génomique et protéomique pour la re-annotation de la base de donnée des séquence génétiques, son application est définie par le séquençage MS qui permet de confirmer la traduction de pseudogènes et d'identifier de nouveaux variant d'épissage ou de nouveaux gènes codant pour des protéines.

L'analyse du contenu protéique d'un échantillon peut être effectuée à travers de nombreuses technologies protéomiques ce qui permet d'obtenir des informations biologiques sous différents angles:

* L'annotation de gènes et l'identification de variants d'épissage: La combinaison des données du séquençage à haut débit de l'ADN ou de l'ARN (DNA-seq ou RNA-seq) avec les données parallèles du séquençage peptidique peut révéler des événement d'épissage alternatifs et peut être utilisé pour l'annotation ou la réannotation du génome.

* L'expression différentielle permet l'évaluation des différences moléculaires entre les types cellulaires tels que les cellules ESCs et iPSCs. Les protéomes peuvent être comparés d'une manière quantitative entre plusieurs échantillons permettant ainsi une analyse différentielle de l'expression protéique.

* L'abondance absolue ou nombre de protéines par cellules permet l'investigation de la relation entre la transcription et la traduction. Lorsqu' effectuée sous des conditions de contrôles, les données protéomiques peuvent être utilisées pour estimer l'abondance absolue des protéines.

* La dynamique temporelle: Les avancées dans le multiplexage ("multiplexing") permet la surveillance temporelle de la dynamique du protéome des processus biologiques complexes.

* La localisation spatiale permet de définir la composition protéique des chromosomes mitotiques. Le fractionnement des organites cellulaires suivie par l'analyse par spectrométrie de masse représente une approche unique pour la description de la localisation protéique. (ESC, cellules souches embryonnaires; iPSC cellules souches pluripotentes induites).

Idéalement, l'intégration de toutes ces approches permettra l'obtention d'une vision claire du protéome 'vivant' avec ses réponses aux signaux extérieurs.

Déchiffrage des régulations post-traductionnelles

Les modifications post-traductionnelles sont des régulateurs clés de l'activité protéique incluant des modifications covalentes réversibles de protéines portant des groupement chimiques, lipides ou d'encore plus petites protéines. Les protéines peuvent être clivées par protéases modifiant parfois la nature chimique des aminoacides. La protéomique basée sur la spectrométrie de masse représente un outil unique pour l'identification et la surveillance quantitative des modifications post traductionnelles globales ou des changement à des région spécifiques (acetylation, phosphorylation et ubiquitylation par exemple). Souvent, plusieurs de ces évènements de régulation coexistent dans une même protéine. L'intégration de ces analyses sert à révéler des mécanismes de 'crosstalk' des modifications post-traductionnelles (PTM) (séquentielle, exclusive ou antagoniste par exemple).

Probablement due à son rôle ubiquitaire dans la plupart des processus biologiques, la phosphorylation reste la modification post traductionnelle la mieux étudiée. Les stratégies d'enrichissement des phosphopeptides, telles que la chromatographie d'affinité (immobilized metal ion affinity chromatography(IMAC)), titanium dioxide (TiO2), chromatographie ou immunoprecipitaion de phosphotyrosine, ont permis de déverrouiller l'analyse de la phosphorylation comme montré par l'étude des différents types tissulaires, des états pathologiques et des lignées cellulaires. Pour comprendre les voies de signalisation, la phosphoproteomique quantitative est utilisée pour surveiller la nature transitoire des évènements globaux de phosphorylation. Comprendre comment les kinases et les phosphatases régulent tous ces sites est aussi important. Les séquences phosphopeptiques ayant lieu in vivo sont identifiée dans les études phosphoprotéomiques à grandes échelles peuvent être analysées par des outils bioinformatiques pour trouver des aminoacides surreprésentés qui flanquent les sites de phosphorylation, fournissant ainsi des indices sur les motifs de séquences linéaires et la spécificité des kinases. Finalement, la génération des anticorps phospho-specifiques des sites candidats permet la localisation des réactions de phosphorylation avec une résolution subcellulaire.

Les expériences d'immunoprecipitation des peptides se font pour l'acétylation et l'ubiquitylation, cependant, bien que l'acétylation des histone est largement appréciée comme étant un mécanisme de régulation des gènes , il a été constaté que l'acétylation cible aussi des milliers de protéines non histone, impliquant un rôle régulateur au-delà du statut de régulation de la chromatine. Le marqueur d'affinité est fusionné au conjugués d'ubiquitine ou de SUMO. Cet assemblage est transfecté dans les cellules d'intérêt , de manière à ce que des cibles de protéine endogène deviennent marqueurs d'affinité durant l'ubiquitylation ou la sumoylation et sont donc isolés par co-immunoprecipitation. L'étude de la N-glycosylation peut être enrichie au niveau des peptides et des protéines en utilisant soit une colonne

d'affinité basée sur la lectine ou la hydrophilic interaction liquid chromatography(HILIC). A la lumière de ces observations, le profil de nombreuses modifications post traductionnelles est déterminé simultanément, incluant la GlcNAc et la phosphorylation. La complexité des modifications post traductionnelles des protéines modifie la compréhension des réseaux de signalisation qui d'un flux d'information linéaire se révèle être un réseau de régulation très complexe et multidirectionnel.

Interactions Protéine–protéine et réseaux biologiques

Les protéines interagissent souvent entre elles dans des complexes multi-protéiques stables ou transitoires de composition distincte. Les protéines peuvent intéragir avec d'autres molécules, telles que les ARN ou les métabolites. Ces complexes ont un rôle essentiel dans les processus régulateurs, les cascades de signalisation et les fonctions cellulaires. Ainsi la perte de la capacité d'interagir peut entraîner une perte de fonction. La caractérisation des interactions protéine-protéine se fait par spectrométrie de masse et purification par affinité. L'analyse de l'interaction protéine-protéine a été initialement dirigé par l'étude double-hybride de levure (yeast two-hybrid (Y2H)) mais a été récemment complémentée par l'utilisation de purification par affinité (AP) de la protéine d'intérêt suivie par MS pour identifier ses partenaires d'interaction. La purification par affinité en tandem (Tandem affinity purification(TAP)) s'est révélée utile dans la cartographie globale des interactions protéine-protéine chez la levure.

La capacité intrinsèque de recombinaison homologue chez la levure introduit directement l'étiquette TAP à un locus endogène choisi, de sorte que les mécanismes naturels de régulation contrôlent l'expression de la protéine de fusion TAP. Une alternative intéressante consiste à utiliser des transgènes clonés en utilisant des chromosomes bactériens artificiels (BAC) et contenant tous des séquences régulatrices endogènes. Les interactions protéine-protéine ont été étudiés au cours de la division cellulaire humaine en combinant un BAC avec une version modifiée de TAP, dans lequel l'une des étiquettes a été remplacée par la GFP, ce qui permet à la fois la localisation et la purification par affinité (LAP). Cela a permis l'observation de protéines s'associant à des composants cellulaires, tels que centrosomes et les spindles, ce qui a conduit à la découverte de plusieurs sous-unités dans des complexes essentiels pour la division cellulaire. L'absence ou la présence de modifications post traductionnelles affecte grandement l'interaction protéine-protéine, comme en témoigne la régulation épigénétique de la transcription par les marques des histones spécifiques et ses «lecteurs». Une méthode très efficace pour identifier ces lecteurs et leurs complexes respectifs a été appliqué dans une lignée cellulaire humaine: les peptides d'histones modifiées ont été utilisés pour précipiter les protéines de liaison , qui ont ensuite été affectés à des complexes de protéines par AP-MS. En plus de l'analyse globale, AP-MS-quantitative peut révéler des informations très pertinentes sur la dynamique des interactions protéine-protéine.

L'identification des "crosslinked sites" peut potentiellement révéler des sites de liaison, et par combinaison avec l'analyse par MS de complexes intact, des informations structurales supplémentaires peuvent être obtenues, comme ce fut récemment montré pour les complexes PP2A. Enfin, ces progrès peuvent être déployés pour démêler les réseaux sous-jacents des maladies, pour lesquelles des mutations de la protéine ou une altération de son expression ou les modification post traductionnelle

compromettent les interactions. Des développements supplémentaires d'outils computationnel sont nécessaires afin de permettre la modélisation du comportement du réseau de protéines dans des conditions changeantes, comme celles déduites des données quantitatives AP-MS.

Applications Cliniques

L'analyse protéomique basée sur la spectrométrie de masse est utilisée pour une étude quantitative approfondie du protéome d'un modèle pathologique et de ses systèmes de contrôles. A l'issue d'une statistique rigoureuse, un ensemble de protéines putatives est défini pouvant être utilisé comme une signature du phénotype. En utilisant des approches plus ciblée, soit basés sur la spectrométrie de masse (selected reaction monitoring (SRM)) soit basée sur des anticorps, ses marqueurs sont validés sur un vaste nombre de patients. Idéalement l'association biologique entre la signature protéique et le phénotype pathologique est corroboré biochimiquement pour confirmer le rôle mécanique du biomarqueur dans la maladie.

Une des applications les plus difficiles de la protéomique est l'identification de biomarqueurs de protéines ayant une valeur pronostique ou diagnostique. Un exemple de biomarqueur réussit qui relie la biologie de la maladie ayant été validé sur un large ensemble de patients est l'inhibiteur de protéase elafin.

Le profil d'expression des échantillons de biopsie est un reflet de la signature moléculaire de la maladie et peut être utilisé pour mettre au point des traitements personnalisés. la protéomique hautement sensitive peut être couplée au laser-capture micro-dissection pour isoler une population cellulaire particulière. Les méthodologies de spectrométrie de masse sont suffisamment avancée pour gérer des échantillons de petites tailles, de manière à ce qu'une couverture raisonnable du protéome peut être obtenue à partir de quelques centaines de cellules. Souvent, les lignées cellulaires in vitro ne reflètes pas les conditions qui existent in vivo. Cette question peut actuellement être résolue par fluorescence-activated cell sorting (FACS) basée sur l'isolation du type cellulaire d'intérêt a partir d'échantillon tissulaires primaires. Ceci permet l'étude de processus biologiques ex-vivo.

Le dérèglement de mécanismes contrôlant les modifications post traductionnelles peuvent avoir de sévères conséquences. Par exemple, des mutation dans les récepteurs tyrosine kinase (RTKs) peuvent induire l'activation constitutive du processus cellulaire régulé. L'immunoprecipitation suivie par la spectrométrie de masse a été utilisé pour identifier les peptides phosphotyrosine dans 41 lignées cellulaires de cancers de poumon et 150 tumeurs. Les signatures phosphotyrosines obtenues ont été utilisée pour classifier les échantillons et ont permis l'identification de kinases activée dans le cancer du poumon, y compris de nouvelle protéines de fusion ALK et ROS. Dans une approche alternative, des expériences d'immunoprecipitation sélective de peptides ont été utilisé sur des substrats phosphorylés des voies de signalisation de la phosphoinositide 3-kinase (PI3K) et la mitogen-activated protein kinase (MAPK). Les biomarqueurs de phosphoproteines découverts par cette nouvelle approche peuvent ensuite être utilisées pour identifier les voies d'activation de PI3K–AKT sans avoir à séquencer les nombreux gènes de la voie mais aussi pour prédire l'efficacité d'inhibiteurs pharmacologiques de AKT,

3-phosphoinositidedependent protein kinase 1 (PDK1), PI3K ou rapamycin (mTOR), démontrant le potentiel de la thérapie personnalisée.

Conclusions et perspectives

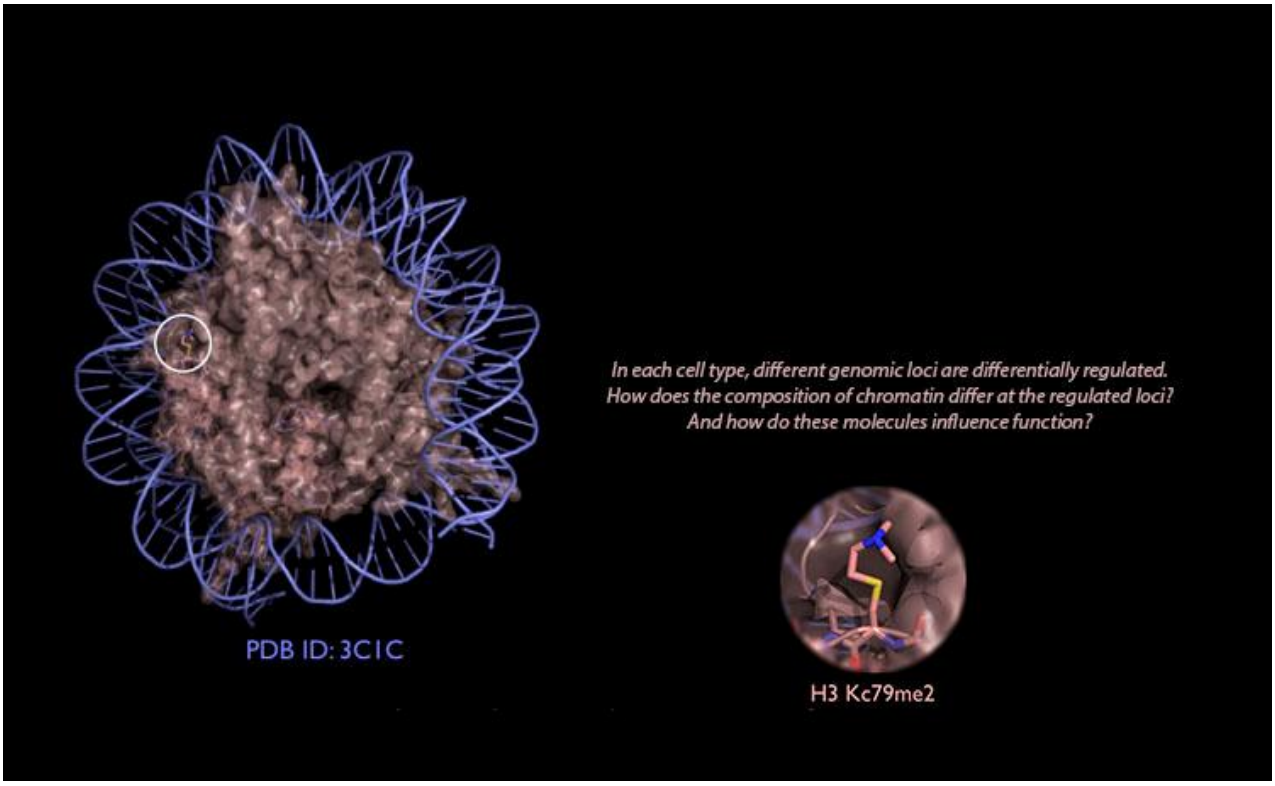
Les futurs progrès dans les technologies de protéomique MS portent sur l'obtention de données protéomiques pertinentes avec un temps d'analyse moindre; la réduction de la quantité de matériau nécessaire permettant une analyse homogènes des populations cellulaires (exemple des cellules triées par FACS) ou des tissus micro-disséqués avec pour but ultime l'analyse du protéome d'une cellule unique. Le développement vers une médecine personnalisée nécessite des snapshots de protéomes personnalisés. Les méthodes basées sur la spectrométrie de masse deviennent plus rapides et compréhensives et l'analyse ciblée des protéines d'intérêt (SRM par exemple) permet de concourir dans un environnement clinique avec des analyses existantes telle que les tests enzyme-linked immunosorbent assay (ELISA).

Dans ce contexte , la spectrométrie de masse qui couple une approche alternative FACS (utilisant des isotopes) avec MS, procure déjà une plateforme hautement multiplexée et compétitive pour des FACS conventionnels. Le défi est d'intégrer davantage de données protéomiques avec des données générées à d'autres niveaux (génome, transcriptome et métabolomes).

L'analyse des données du protéome ont été effectué en comparaisons aux données du séquençage d'ADN et d'ARN et la comparaison des données protéomiques avec celle obtenues du profilage ribosomique. Ces méthodes des séquençage d'ADN et d'ARN en développement sont probablement plus complets pour la couverture du génome entier que pour le profilage du protéome sans avoir pour autant la possibilité d'interroger certain niveaux essentiels de la régulation de l'expression génétique et la biologie des réseaux (modifications post traductionnelles par exemple). Les approches de la biologie intégrative sont donc essentielles pour aborder les questions biologiques à l'échelle du système.

Résumé de l'article 2

Job Dekker, Marc A. Marti-Renom, Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics.* mai 09, 2013, pp. 390–403.



http://simonlab.commons.yale.edu/

Exploration de l'organisation tridimensionnelle des génomes : interprétation des données de l'interaction chromatinienne

L'intersection des territoires chromosomiques entraine un brassage génétique mêlant des interactions fonctionnelles entre les loci de différentes régions chromosomiques. Les gènes activement transcrits ont tendance à être co-localisés dans des groupes spécifiques en relation avec l'élément de régulation de la transcription. En effet, la transcription à lieu dans des sites subnucleaires enrichis en ARN polymérase II et autres facteurs transcriptionnels. Les segments inactifs du génome ont tendance à s'associer à la périphérie nucléaire ce qui dénote un caractère spatial et fonctionnel de la compartimentation nucléaire où le positionnement subnuclaire du loci est corrélé à l'expression génétique.

Cet article décrit les différentes approches statistiques et computationnelles mises au point pour analyser des données de l'interaction chromatinienne telles que les méthodes basées sur la capture de la conformation chromosomique (3C) et leurs applications dans la mise en évidence de l'organisation spatiale des chromosomes, les nouveaux aspects de la structure chromatinienne en corrélation avec l'expression génétique et sa régulation.

Méthodes basées sur 3C

Les méthodes basées sur la capture de la conformation chromosomique (3C) permettent la détermination de la fréquence pour laquelle chaque paire de loci dans le génome se trouve dans une proximité physique suffisamment proche (intervalle 10-100nm).

L'analyse des bases de données en C3 se fait par trois approches:

- La première approche a pour but d'identifier les loci interagissant fréquemment, la formation des boucles de chromatine ou les événements de co-localisation spécifique. L'analyse des interactions spécifiques des loci a été utilisée pour déterminer la complexité du domaine chromosomique.

- Les deux autres approches- " restraint-based modelling" et des approches modélisant la chromatine comme polymère utilisent des données d'interaction de base et non spécifiques pour la construction de modèles spatiaux des chromosomes. Les modèles 3D peuvent être utilisés pour identifier des éléments structuraux complexes et des éléments d'ADN impliqués dans l'organisation chromosomique. Les modèles 3D servent aussi dans l'estimation de la dynamique chromatinienne et la variabilité intercellulaire des repliements.

Les approches basées sur 3C permettent la cartographie du repliement chromosomique avec une bonne résolution pour une observation des gènes et des éléments de régulation à l'échelle du génome.

Dans des expériences classiques de 3C, les produits de liaisons simples sont détectés par PCR en utilisant des amorces spécifiques aux loci. La méthode 4C (3C' circulaire ou '3C- on-chip') utilise une PCR inverse pour générer des profils d'interaction génomique pour un locus. 5C combine 3C avec des approches de capture hybride afin d'identifier en parallèle des millions d'interactions ayant lieu entre deux grands ensembles de loci (entre un ensemble de promoteurs et un ensemble d'éléments régulateurs distaux par exemple). Les approches 4C s'appliquent à tout le génome mais se limitent à un seul locus. Les analyses 5C font intervenir plusieurs milliers de fragments de restriction ce qui met en évidence des millions d'interactions sur des dizaines de megabases pouvant être contigües ou dispersées dans le génome parmi les loci d'intérêt.

La méthode Hi-C fut la première adaptation non biaisée de 3C couvrant tout le génome. Cette méthode fait intervenir une étape de remplissage par nucléotides biotinylés des extrémités d'ADN après une digestion par des enzymes de restriction. La Hi-C procure une cartographie complète des interactions mais la résolution de cette cartographie est dépendante de la profondeur du séquençage.

3C et 4C génèrent des profils d'interaction pour des loci uniques. Les technologies basées sur la capture de la conformation des chromosomes (3C) détecte les loci en proximité spatiale ceci étant démontré à travers des exemples biologiques de structures particulières.

Les méthodes 5C et Hi-C ne se limitent pas à un seul locus d'intérêt mais génèrent des matrices de fréquences d'interactions qui peuvent être représentées comme un "heat map" bidimensionnelles avec des positions génomiques le long des axes.

Interprétation des données de l'interaction chromatinienne

Les essais basés sur 3C montrent une fréquence relative de la proximité spatiale de deux loci dans la population cellulaire. Cependant ces essais ne permettent pas de distinguer des associations fonctionnelles et non fonctionnelles et ne révèle pas les mécanismes de leurs co-localisation. La proximité spatiale peut être le résultat de contacts directs et spécifiques entre deux loci assurée par des complexes protéiques de liaison ou peut résulter de la co-localisation indirecte des pairs de loci dans une même structure subucléaire, telle que la lamina nucléaire, le nucléole ou la machinerie transcriptionnelle. La co-localisation cellulaire peut être due au repliement de la fibre chromatinienne par interactions spécifiques, d'autres contraintes, ou due à des collisions aléatoires (non spécifiques) dans un noyau encombré.

Liaison des éléments régulateurs aux gènes cibles

La mise en place des boucles de chromatine permet la communication des éléments régulateurs avec des gènes cibles apparentés réunissant en proximité spatiale des éléments espacés dans un génome linéaire.

La méthode 3C est utilisée pour quantifier les fréquences d'interaction entre un élément d'intérêt (un promoteur par exemple) et les régions flanquantes de la chromatine qui s'étendent jusqu'à des centaines de kilobases. L'analyse de tel profils d'interaction permet la mise en évidence de loci distaux interagissant plus fréquemment avec des loci d'ancrage ce qui dénote des interactions en forme de boucle. Il fut également démontré que les fréquences d'interaction diminuent exponentiellement en fonction de l'augmentation de la distance génomique.

Analyse des boucles de chromatine

La méthode 5C permet une analyse complète des boucles de chromatine pour un grand nombre de gènes mesurant en parallèle plusieurs profils d'interactions. La présence des interactions en forme de boucle est démontrée lorsqu'une paire de loci interagit plus fréquemment statistiquement avec une valeur supérieure à la fréquence de base.

Plusieurs événement de formation des boucles représentent des interactions cellulaires spécifiques entre promoteurs de gènes actifs et des éléments distaux tels que les "enhancers" actifs. Ceci est en accord avec le rôle de ces structures chromosomiques dans l'activation des gènes. Certaine de ces interactions distales incluent des promoteurs en boucles avec les sites de liaison par la protéine CTCF.

Les éléments régulateurs sont souvent considéré contrôlant le gène le plus proche. Cependant le profil moyen des interactions qui mènent à la formation des boucles autour des promoteur est asymétrique: Les promoteurs interagissent avec des éléments distaux localisés en amont ou en aval du site d'initiation de la transcription (TSS), mais les interactions en boucles les plus fréquentes sont observés sur des éléments localisés à 120 kb en amont du site d'initiation de la transcription (TSS).

Topologically associating domains

Les chromosomes sont composés de TAD ou "Topologically associating domains" de centaines de kilobases. Les loci situés dans les TADs interagissent entre eux avec une fréquence plus grande que les loci se trouvant à l'exterieur du domaine. La mise en évidences des TADs a été permise par analyse des profils de cartographie Hi-C à faible résolution des génomes humain et de souris en combinaison avec des approches utilisant les modèles de Markov cachées (HMMs). Cette analyse a permis de montrer le rôle des TADs comme composants fondamentaux et universels des chromosomes avec pour rôle principal de limiter les gènes qu'à certains éléments régulateurs de l'expression génétique. La présence des TADs explique la corrélation en expression des groupes de gènes voisins de la chromatine structurale pour tous les types cellulaires.

Construction d'un modèle tridimensionnel de la chromatine

Les représentations 3D permettent l'identification de caractéristiques plus complexes de la conformation des chromosomes telle que la formation de domaines globulaires, des territoires chromosomiques, l'identification de la séquence d'éléments et des mécanismes impliqués dans le repliement.

Les approches de modélisation 3D peuvent être divisées en deux types de méthodes. Dans la première approche, un ensemble de données d'interaction de la chromatine est utilisé pour obtenir une conformation 3D de la moyenne des populations. La deuxième approche couvre les donnés d'interaction de la chromatine qui sont analysés en termes statistiques de l'ensemble des polymères.

Les cartes d'interaction complètes reflètent la fréquence de co-localisation des loci dans la moyenne des population qui est inversement proportionnelle à la distance spatiale moyenne. La fréquence d'interaction déduite est utilisée comme contrainte pour la construction de modèle tridimensionnels plaçant les loci en espace 3D relatif de manière cohérente avec leurs probabilités d'interaction.

Les mécanismes itératifs et intégratifs pour la construction des modèles

Le mécanisme itératif consiste à l'acquisition des données, la représentation, l'optimisation et l'analyse des modèles ainsi qu'à l'établissement des scores.

L'implémentation de la modélisation 3D basée sur la contrainte du domaine génomique fut principalement celle de l'analyse spatiale du locus de l'immoglobuline humaine H (IGH) utilisant des mesures de distance obtenues par les imageries optique d'un ensemble de 12 positions dans le locus.

Le développement dans la construction des modèles 3D permettra de définir les différents niveau de l'organisation chromosomique (y compris les mécanisme de formation de boucles, les globules ou les TADs) afin de localiser avec précision des éléments de séquence impliquées dans ces structures et de placer les loci espacés dans un contexte spatial révélant des relation potentiellement fonctionnelles et de longue durée.

Les approches de polymères

Permettent de mettre en évidence les caractéristiques de l'organisation statistique des états de repliement des chromosomes, leur variabilités cellulaires et leur dynamique intracellulaire. D'autres études ont mis en oeuvre des simulations d'anneaux de polymères pour mettre en évidence des territoires chromosomiques formés par contraintes topologiques prévenant le brassage des chromosomes individuels. Les simulations de polymères sont un outil d'investigation pour montrer l'influence des propriétés de la fibre chromatinienne, son repliement local et les interactions inter chromosomiques spécifiques sur la localisation des chromosomes.

Les modèles physiques de polymère sont utilisées afin de mesurer les probabilités des interactions spatiales. Les fréquences de contact mesurées sont utilisées pour caractériser l'ensemble des conformations de la chromatine.

L'absence de contacts reproductibles sur de grandes échelles fait que la conformation complexe des chromosomes soit très différente des conformations d'une seule protéine repliée. Ceci suggère que la chromatine à de large échelles (>1Mb) peut être mieux caractérisée en tant qu'ensemble statistique de diverses conformations, probablement reflétant des différences intracellulaires avec des propriétés statistiquement spécifiques, spatiales ou topologiques.

Les interactions ayant lieu dans un bras chromosomique montre une décroissance frappante de 100-repli de la probabilité de contacte P avec la distance génomique s, ce qui en fait la caractéristique la plus importante des interactions intrachromosomiques. Les données Hi-C pour une cellule humaine non synchronisée mettent en évidence trois régimes qui chacun exhibent une décroissance dans la probabilité de contact.

Le globule fractal résultant de la condensation du polymère dont les contraintes topologiques permettent d'éviter les noeuds et ralentissent l'équilibration du polymère présente certaines propriétés particulières:

- une condensation de la chromatine d'une manière dense et uniforme à l'échelle de <10Mb est compatible avec les globules de chromatine observés d'à peu près 1 µm de diamètre.

-La conformation sans noeuds des globules fractaux (qui n'est pas une caractéristique des globules d'équilibre) permet une ouverture et une fermeture facile ou une translocation de régions chromosomiques sur de larges distances dans le noyau.

- La compaction dense des segments de globules fractaux implique une région continue du génome et non dispersée. Cette propriété distingue le globule fractal du globule d'équilibre dans lequel des segments individuels s'apparentent à des marches aléatoires (étendus et mélangés).

Perspectives futures

Les améliorations dans l'analyse des données expérimentales et computationnelles permettra de faciliter l'adressage de plusieurs question importantes concernant le domaine de régulation du génome.

Plusieurs études basées sur la méthode 3C ne permettent pas la mesure directe de la dynamique des variations intercellulaires du repliement chromosomique. Ceci entraine des incertitudes sur la stabilité intracellulaire des interactions formant des boucles de chromatine et sur leur comportement aléatoires et stochastiques entre les cellules.

La contribution relative de la séquence génomique et de l'activité transcriptionnelle dans l'établissement de l'architecture en compartiment des chromosomes demeure inconnue aussi bien que le rôle de l'association des lamines, les co-localisations directes ou indirectes des régions transcrites et autres mécanismes moléculaires façonnant l'organisation associée à l'activité du noyau.

Le développement rapide des technologique dans ce domaine portera la lumière sur la structure chromosomique variable durant le développement, en réponse aux perturbations et sur le repliement et le changement de conformation chromosomique durant le cycle cellulaire.